The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010.

Jack A. Gilbert[1,2], Folker Meyer[1,3], Janet Jansson[4], Jeff Gordon[5], Norman Pace[6], James Tiedje[7], Ruth Ley[8], Noah Fierer[6], Dawn Field[9], Nikos Kyrpides[10], Frank-Oliver Glockner[11], Hans-Peter Klenk[12], K. Eric Wommack[13], Elizabeth Glass[1], Kathryn Docherty[14], Rachel Gallery[14], Rick Stevens[1], Rob Knight[6]

[1]Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, U.S.A.
[2] Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.
[3] Computation Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.,
[4] Lawrence Berkeley National Laboratory • Earth Sciences Division 1 Cyclotron Rd., MS 90-1116 • Berkeley, CA 94720
[5] Center for Genome Sciences & Systems Biology, Campus Box 8510, 4444 Forest Park Blvd., Room 5401, St. Louis, MO 63108, USA.
[6] Department of Chemistry and Biochemistry, UCB 215, Boulder, CO 80309-021
[7] 540 Plant and Soil Sciences Building, Michigan State University, East Lansing, MI 48824-1325, USA.
[8] Department of Microbiology, 260A Wing Hall, Cornell University, Ithaca, NY 14853, USA.
[9] NERC Centre for Ecology & Hydrology, Crowmarsh Gifford, Wallingford Oxford, OX10 8BB, UK.
[10] DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
[11] MPI for Marine Microbiology, Celsiusstr. 1, D-28359 Bremen, Germany
[12] DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Inhoffenstraße 7 B, 38124 Braunschweig, Germany.
[13] University of Delaware, Delaware Biotechnology Institute, 15 Innovation Way, Newark, DE 19711, USA.
[14] NEON, 1685 38th St., Suite 100, Boulder, CO 80301, USA.

Abstract

This report details the outcome from the Earth Microbiome Project's first meeting to discuss sample selection and acquisition. This meeting was held at the Argonne National Laboratory on Wednesday October 6th 2010 between 2pm-5pm. The meetings aim was to discuss how to identify environmental samples to be sequenced and analyzed using metagenomics as part of the EMPs global effort to systematically determine the functional and phylogenetic diversity of microbial communities across the world.

Introduction

To understand microbes (Bacterial, Archaeal, Eukaryal and Viral) in terms of whom they are and what they do is the challenge of microbial ecology. The Earth microbiome Project presents a revolution in how we tackle this problem and defines both questions

and potential suite of tools to provide answers. The EMP will provide a quantum leap in our ability to interrogate this community through a truly global collaborative project. Approximately 1 x 10^{30} microbial cells live on this planet at any one time. To date, the total global environmental DNA sequencing effort (metagenomics) constitutes significantly less than 1 percent of the total DNA found in a liter of seawater or a gram of soil. Hence, we have vastly under sampled the complexity and diversity microbes on this planet. However, recent advances in high-throughput sequencing technologies have provided an unprecedented opportunity to explore the microbial universe, and we propose to leverage this capability at a scale many orders of magnitude greater than any previously conceived study. We wish to sequence microbes and microbial communities from every conceivable biome. By exploring genetic information from every ecosystem we hope to achieve 3 main goals:

1. To explore the protein universe and attempt to produce complete an inventory of protein family diversity.
2. To define the microbial community structure and explore at different scales what structures it, i.e. defining microbes in environmental parameter space – a microbe centric view.
3. To create a global database of samples, genes and proteins the can be used to answer fundamental questions about the ecology of life on and off the earth.

Microbes are the life-support system for our planet. Without them there could be no other life—and yet we know virtually nothing about how they provide this support. We as a species are now having a significant impact on this planet, for example, changing the weather systems and biogeochemistry of the atmosphere and acidifying the oceans. These large-scale changes will affect microbial life and, with it, all life on Earth. It is therefore essential that we, the scientific community, develop a strategy to improve our understanding of the role and importance of microbes.

The EMP aims to select and acquire between 100 and 200 thousand samples from numerous and diverse environments across the world. Therefore the bottleneck for this project will be identifying projects which can provide samples, determining whether the samples adhere to the stick requirements for associated metadata per sample that we want to implement, and the infrastructure, protocol and legal implications of such an endeavor.

This was a closed meeting with 16 in-room participants and 2 on-phone participants. The format was a discussion forum. Therefore this report will be divided into sections based on topics of discussion detailing the key output of that discussion.

What do we want from samples?

Prior to discussion regarding what types of samples and what we wanted to get from them, Jeff Gordon suggested and it was agreed that it was necessary to identify the principal stakeholders. These were primarily identified at carbon-cycle and climate researchers, agricultural and human health organizations, fundamental science through the National Science Foundation (NSF), ecological & biodiversity research and interest

groups (for which products must be developed as deliverables for the community). Rick Stevens also stipulated that there was potential commercial interest in novel functions or greater enzyme efficiency, which could be fed by data from this study.

Jeff Gordon also suggested that samples for the EMP should come from sites or locals that capture the public imagination. For example, World Heritage Sites and sites with obvious desecration, also Super Fund sites would be ideal targets. Especially for the polluted or desecrated sites it was very important that we determine whether the environment can heal itself or whether we need to identify roles for microbes in attempting to provide an anthropogenic solution to pollution. It was agreed that it is very important to frame the selection of sites. The group went on to discuss how to capture peoples imagination in a similar way in which space does. Jeff Gordon stipulated we need a grand challenge statement, as when President Kennedy said "Let's send a man to the moon" in 1961. Ruth Ley highlighted that in order to excite people we needed to design extraordinary visuals. Jack Gilbert suggested that it was essential to 'show' microbial communities to the public.

Norman Pace said that it is very important that the EMP sample acquisition does not constitute 'just-another-survey'. While Rick Stevens said that there is still huge interest in dark matter, which by definition needs to be explored. The global microbiome and its diversity is one of the most comprehensive dark-matter problems, because huge amounts of biodiversity have not been explored. It was suggested that we have to be more quantitative about how we proceed. It may even be necessary to make some controversies about alternative projections. Physics is never short of making predictions about what is true and then having it confirmed. Are there conjectures we can make about diversity and then get this resolved through the scientific process? Noah Fierer suggested that physicists were much better organized, a group behavior we need to foster in the biological community. Noah went on to stipulate that microbiologists have an option to explore communities on the basis of pathogenicity and virulence, which could be used to excite the imagination.

Jack Gilbert went on to describe how to potentially select samples. It is necessary that the samples enable the production of a 'topographical' map of microbial function – which will be the most useful deliverable for the benefit of mankind. Rick Stevens stipulated that we therefore can use a phylogenetic survey to explore diversity and composition prior to targeted metagenomic (functional) discovery. Nikos Kyrpides suggested that maybe a phylogenetic profile was all that was necessary at this stage as we don't even know the distribution of microbes and this could be a primary goal in the short term.

What data do we want associated with these samples?

It is essential that we have high quality environmental contextual data associated with every sample. Therefore we need a mechanism by which to 'grade' samples by the quality of their metadata. James Tiedje indicated that most soils samples collected to date have very poor metadata, although this was improving. Rob Knight indicated that through the work of organizations like the Genomic Standards Consortium (www.gensc.org)

more samples would be collected with better metadata in the future. Janet Janssen suggested that the issue was standardization and quality assessment of metadata as well as sample quality. Janet pointed out that it is essential that we only collect high quality samples with high quality metadata. Norman Pace indicated that it is very important to get the chemistry of a sample collected. James Tiedje suggested that it was also extremely important to have information regarding the sample processing, e.g. soil researchers often only have air-dried samples, and often very little is known about the impact of these methods on downstream molecular analysis. Rob Knight argued that a role of the EMP would be to fund experiments to demystify the superstition regarding the impact of different sample preparation and experimental procedure on bias in the community analysis.

It will be absolutely necessary to have environmental parameters with all samples so that we can reduce the redundancy in sample analysis and explore a greater diversity of biomes for financial investment. Rick Stevens asked if there was a way to articulate classes of environments, i.e. parameter space for soils, marine ecosystems, lakes and rivers, etc. This information will help us to design a systematic way to sample environments, to enable a systematic march through the ways in which microbes live. Rachel Gallery indicated that NEON (http://www.neoninc.org/) chose different ecosystem types, including freshwater, to explore a diverse array of environments for long-term monitoring. They identified these by continental eco-regions. Environmental nomenclature was suggested as a method for choosing ecosystems, it was indicated that the ontology community had a vast range of tools to implement this, e.g. Habitat Lite (http://gensc.org/gc_wiki/index.php/Habitat-Lite) for microbial ecosystems.

Reaching out to the Scientific Community?

Jack Gilbert pointed out that we need to come up with an efficient model for sample acquisition, and went on to suggest that the EMP publish a short letter in every relevant journal highlighting the fact that we are looking for samples. Additionally, through advisory board members and involved parties we can 'reach-out' to the community through colleagues to identify excellent sample datasets. Rob Knight commented that the scientific community must drive sample collection, for example Margaret McFall has a list of species for host-associated samples. It is vital that we reach out to all microbial ecologists who have already collected samples that would have good metadata. Noah Feirer identified that Texas A&M have been collecting cow fecal samples, which is now at approximately 10,000 samples, and have excellent metadata. All these samples are frozen. Rachel Gallery suggested that the plant research community had many thousands of samples regarding plat pathology and endosymbiosis, which could be targeted. Jack Gilbert pointed out that Oliver Ryder had identified a wealth of host-associated samples from zoo animals in San Diego, USA. Jim Tiedje defined one model, which would be to take all samples from all groups that meet the standards, from any ecosystems. This could represent a first pilot study, and could be done very soon. Rob Knight pointed out that this is very like the Community Sequencing Project from the Joint Genome Institute and we could implement that model. Norman Pace also agreed that this would be the most effective way of implementing a rapid development of the EMP.

Rick Stevens went on to discuss the importance of choosing an effective sequencing strategy for sample acquisition. He suggested that in order to verify samples from different researchers we would need to provide a standard sample per sequence run or DNA extraction. Noah Fierer suggested that we should select projects with no fewer than 100 samples; otherwise the economics would not work. Folker Meyer suggested that we should choose samples from ecosystems and biomes that have not yet been analyzed using this technology. It would be up to the advisory board to determine which samples should be analyzed.

Overall is was agreed that the board represents a broad range of communities and that each person should 'reach-out' to their community and identify the types of projects that are available or will become available. This information should then be added to a central repository called the Global Environmental Sample Database (GESD) which will be used to grade, refine and select the environments for a series of pilot studies and subsequent analyses.

How do we collect the samples?

Sample collection or acquisition was identified as the biggest problem. Rob Knight suggested that we need an infrastructure to collect samples and to insure good quality samples and associated metadata. He went on to point out that it is vital that we recruit people to the EMP who are excited about contributing quality samples. However, the focus should first be on hypotheses testing, answering the EMP central tenant questions. Jack Gilbert suggested that we should consider whether to accept samples or possible just DNA. Rob Knight identified that if we are to select DNA we need some standards for DNA quality. Additionally, we should consider sending primers to determine whether the DNA is amplifiable prior to sending to the EMP. Folker Meyer suggested that we should maybe only select projects that have demonstrated the ability to get good sequence from Illumina platform from the samples they are sending. It was generally agreed that at this stage this would be overly difficult for the majority of research groups. Rick suggested that it would be essential to implement a standard sampling protocol for all environments now, one for each would be practical. It was agreed that this would be ideal but also very difficult, requiring a consortium for each community to come together and implement and enforce these standards. James Tiedje pointed out that NEON is an ideal example of these standards, but that this was not necessarily the biggest problem. Obtaining samples from outside of the US would be the most significant problem, licences and permits would be required, and the countries from which samples were sent would need to agree to sequencing and downstream analysis to prevent litigation. Rick Stevens suggested that the EMP could potentially ship a sequencer in to the country and this would prevent shipping costs and permits for the physical samples or DNA. Rob Knight suggested that one possible solution would be to have visitors come and extract samples at an EMP affiliated Laboratory.

Ownership of samples and data?

James Tiedje pointed out that probably half he people will not participate in the EMP because they would fear that they will lose control of their own samples and data. Rob Knight pointed out that the EMP needs to educate people to the fact that the EMP as a network will allow researchers to do more than trying to analyze the data on your own. Janet Janssen suggested that people will be amenable if they are guaranteed to have publication rights. Jack Gilbert suggested that we need to get over the 'bio-ego' that is pervaisive in our community. It is essential that this is done in an open and collaborative way, and by doing so they will have access to a more comparable and complete dataset than ever before. James Tiedje agreed that the ego was a big problem, and possible insurmountable; if the EMP is successful as a pilot project and it can prove that the data generated is more effective than data in isolation, then the community will be more receptive. Frank-Oliver Glockner suggested that people may still not buy into the bigger vision, this is an issue of trust and people will fear the autocracy of the EMP. James Tiedje sais that in other disciplines this is not such a problem, for example the human genome publication all met together and participated in project with 125 authors. Jack Gilbert pointed out that people tend to trust the 'village elders' of a community, and if they support the EMP publically then people will follow.

Moving forward?

Nikos Kyrpides suggested that we need to work towards a NASA style institute that binds us all together under the EMP umbrella. However, in the short term, the EMP requires a pilot study to demonstrate the benefit of this collaborative and comparable research initiative. Jack Gilbert suggested that the EMP should use a 16S ribosomal RNA gene survey to produce a map of 100,000 samples, and then select a range of samples for ultra-deep sequencing. Jack also noted that Katrina McMahon has already signed on and provided an extensive series of temporal and biogeographic samples from temperate lakes, additionally Argonne National Laboratory also has a total of ~8000 samples ready to go for a broad 16S rRNA gene study which will be used to target metagenomics. Rob Knight asked the board to start making a compilation of other sample collections which they could gain access too. Janet Janssen pointed out that a large number of samples had already been sequenced and another approach would be to start compiling comparable metagenomic and 16S rRNA gene datasets to show the value of a centralized comparable network. Rob Knight and Folker Meyer pointed out that we need to make the data and analysis freely available to everyone. This must be as open as possible and contain a significant educational element. Dawn Field and Janet Janssen suggested that school reach out programs would be very effective. Rob Knight summarized the meeting by stipulating that we will try to demonstrate the value of the EMP initially with in-house samples. The board will reach out to networks for existing samples, such as Terragenome, Tara-Oceans and NEON, and solicit additional samples through professional contacts. The EMP should also work on a special Science Issue sponsored by Illumina and MoBio through which we advertise the EMP and promote community integration.

Acknowledgements

We would like to thank Argonne National Laboratory for hosting the meeting and Darlyn Mishur for organisation. This work was supported in part by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357.